УДК 681.3

А. Г. Додонов[1], Д. В. Ландэ[2]
[1]Институт проблем регистрации информации НАН Украины
ул. Н. Шпака, 2, 03113 Киев, Украина
[2]Информационный центр «ЭЛВИСТИ»
ул. М. Кривоноса, 2а, 03037 Киев, Украина

Организация сети информационных прокси-серверов

Приведены проблемы современного web-пространства, не позволяющие рассматривать его как надежный и живучий информационный ресурс. Описан подход к организации сетевой инфраструктуры, позволяющей решить ряд проблем на основе использования системы контент-мониторинга и сети информационных прокси-серверов.

Ключевые слова: *Интернет-ресурсы, информационный прокси-сервер, надежность, доступность, сетевая инфраструктура, контент.*

Возможности доступа к Интернет-ресурсам, которые привлекают своей открытостью, объемами и содержательной многогранностью на первый взгляд кажутся безграничными. Однако кризисные события в разных областях, будь-то крупные теракты или чемпионаты по футболу, свидетельствуют об обратном. Именно в кризисных ситуациях Интернет достаточно часто подводит. Существует множество проблем — от перегруженности сетевой инфраструктуры до вирусных атак, уязвимостей и отказов в обслуживании отдельных web-серверов. Целый ряд проблем порожден также объемами, разнообразием представления и динамикой контентной части сетевых информационных потоков.

Проблемы

Несмотря на такие позитивные качества как открытость и доступность, существующую инфраструктуру Интернет нельзя признать надежной, живучей и достоверной [1]. Назовем еще несколько проблем, присущих современному web-пространству.

1. Не решена задача доступа пользователей к разнородным web-ресурсам «из одного окна» для получения обобщенного представления потоков информации по необходимой тематике.

2. Не обеспечена возможность своевременного «напоминания» и «проталки-

вания» профильной для пользователя информации, публикуемой на большом количестве web-сайтов.

3. Достаточно высокая вероятность отказа в обслуживании со стороны критически важных Интернет-ресурсов в самое необходимое время.

Известно, что сегодня существуют технологии интеграции контента, частично предоставляющие решение названных проблем, однако не исследован уровень безопасности их применения, возможно массового. Вопросы сетевой безопасности, например, в рамках современной концепции Семантического Web, по мнению авторов, выглядят преимущественно декларативно, а на практике заужены тематикой цифровой подписи.

Из всего сказанного выше следует необходимость создания новой инфраструктуры, обеспечивающей надежную доставку сетевого контента заинтересованным лицам и организациям, в частности, на государственном уровне.

Ограничения, с которыми необходимо считаться

Пожалуй, самая распространенная причина отказов от предоставления web-сайтами своего контента по запросам пользователей состоит в их банальной перегруженности. Вместе с тем мало кто из информационных администраторов web-сайтов, даже сайтов и порталов органов государственной власти, владеют данными о максимально возможном количестве запросов пользователей, которые способны удовлетворить эти ресурсы. Владельцы любительских web-сайтов и сайтов электронных СМИ даже не задумываются об этом вопросе.

При этом существуют достаточно жесткие ограничения возможностей web-сайтов при массовой работе с их контентом. Следует заметить, что многие из этих ограничений не учтены даже в нормативных документах, регламентирующих требования по защите информации на web-страницах [2]. Назовем некоторые из них, которые влияют на уровень доступности web-ресурсов:

— ширина канала связи до web-сайта. Это ограничение было наиболее обосновано на начальных этапах развития сети Интернет;

— физические ограничения программно-технических платформ web-серверов. Для снятия этого ограничения, например, популярные поисковые службы используют сотни Frontend-серверов;

— устанавливаемые ограничения в программном обеспечении web-серверов. Например, у самого популярного в настоящее время web-сервера Apache [3] параметром MaxKeepAliveRequests определяется максимальное количество разрешенных запросов при устойчивом соединении. При этом для обеспечения максимальной производительности это значение зачастую устанавливается по умолчанию равным 100;

— ограничения на отдачу динамических страниц, например, со стороны СУБД, поисковых систем или сервисных других программ. Такие ограничения часто устанавливаются при совместном виртуальном хостинге у провайдеров и измеряются количеством запросов в час. В случае использования популярной в Интернет СУБД MySQL [4] это ограничение, например, задается параметром max_questions, значение которого, как правило, составляет 72000 (20 обращений к базе данных в секунду). Превышение ограничения может происходить по разным

причинам: установка малого значения в соответствии с политикой провайдера, высокая посещаемость сайта, установка ресурсоемких приложений типа статистики, нестандартных программ и т.д.

Следует выделить два явления, которые существенно влияют на надежность получения информации от web-сайтов: пиковые нагрузки со стороны пользователей в кризисные дни (например, 11 сентября, «Оранжевая революция», начало войны в Ираке и т.п.) [5] и DoS-атаки (Denial of Service или Отказ от обслуживания). Во втором случае хакеры особым образом формируют запросы к программным компонентам web-серверов, чтобы загрузить их до такого уровня, когда они перестанут функционировать. Такие атаки, как правило, не ведут к разрушению самих серверов; чтобы вернуть web-сервер в рабочее состояние, как правило, требуется перезагрузка. Часто DoS-атака выполняется с большого количества компьютеров, в этом случае она называется распределенной (DDoS Distributed Denial of Service). Этот вид атак можно отнести к так называемым «сетевым войнам», формам организации конфликтных ситуаций на основе Интернет [6]. В таких случаях web-серверы не успевают отвечать на все запросы, в том числе и запросы реальных пользователей.

Обе ситуации — и злонамеренная DoS-атака, и кризисная пиковая посещаемость приводят к недоступности информационных ресурсов web-сайтов, в частности, для аналитиков и лиц, принимающих решения.

Поведение систем в результате возникновения данных ситуаций: определенное количество запросов может обрабатываться — остальные стоят в очереди или «отбрасываются» по тайм-ауту.

Назначение прокси-сервера

Как подход к решению названных проблем предлагается построение сети — системы связанных информационных прокси-серверов. Необходимо заметить, что использование прокси-серверов (точнее, кэширующих прокси-серверов) при работе в сети Интернет очень популярно [7]. В этом случае прокси-серверы служат, в основном, для ускорения загрузки страниц за счет кэширования содержимого страниц, ответов на запросы пользователей, DNS и т.п.

Для английского слова «proxy» в данном контексте применимы такие переводы: «полномочный представитель», «посредник». В Интернет-технологиях прокси — это программа, которая получает запросы, обращается к внешнему сервису из Интернет, получает ответы и возвращает их пользователям. Под кэшем понимается информационное хранилище, в котором хранятся часто запрашиваемые web-страницы.

Именно идеологию кэширующего прокси-сервера предлагается рассмотреть как базу для построения инфраструктуры, которая позволит решить проблемы, названные в статье.

При этом к данным, которые предположительно будет обслуживать информационный прокси-сервер, предъявляются такие требования:

— рассматривается динамическая новостная составляющая web-пространства как наиболее критичная с точки зрения обеспечения оперативного доступа;

— множество кэшируемых web-сайтов выбирается экспертами в соответствии с их вкладом этих источников в информационное пространство и может ограничиваться несколькими тысячами;

— информация в прокси-сервере должна быть представлена в универсальном внутрисистемном формате, предполагающем однозначную синтаксическую трактовку. Этим форматом может быть популярный сегодня XML или один из его диалектов (например, RSS);

— данные в информационном хранилище (кэше) должны обновляться и ротироваться по расписанию, соответствующему динамике их обновления на web-сайтах.

Прокси-сервер, с одной стороны, предназначен для надежного обслуживания пользователей корпоративных сетей, а с другой стороны, может обеспечивать обмен данными с аналогичными внешними прокси-серверами. Такое взаимодействие образует своеобразную сетевую структуру, которая, по мнению авторов, может оказаться решением названных проблем.

Принципы функционирования информационного прокси-сервера

Пользователи информационного прокси-сервера обращаются к данным, помещаемым в информационное хранилище (кэш). Кэш пополняется программой-роботом, которая сканирует целевые web-сайты. Следует отметить, что многие популярные сетевые информационно-поисковые системы также кэшируют информацию с web-страниц, предоставляя ее при необходимости пользователям. Можно назвать такие системы, как Yandex (режим «Сохраненная копия»), Rambler (режим «Восстановить текст»), Gooogle (режим Cached).

Характерная особенность роботов — настойчивость (при получении отказов на запросы, он продолжает их задавать до момента получения позитивного ответа). Это тот плюс, который, например, позволил авторам наблюдать поток сообщений из Вашингтона 11 сентября при общем впечатлении об Интернет, как «зависшей» в тот момент сети.

Интеллектуальный сканер системы (рис. 1) обращается к web-сайтам и скачивает с них информацию по сценарию, составленному на специальном языке макроописаний [8]. При этом сценарии могут существенно отличаться по качеству, все зависит от квалификации эксперта-оператора.

Предполагается, что в результате сбора и первичной обработки данные в информационном хранилище будут программно приведены к единому формату, классифицированы в соответствии с определенными рубрикаторами, каждому документу приписан ряд дескрипторов, включая ключевые слова.

Вместе с тем администраторам web-сайтов известны многие роботы, которые излишне загружают их ресурсы, не принося при этом явной пользы. Опасность массового применения роботов состоит в том, что они сами могут порождать нечто подобное DoS-атакам. Что можно противопоставить этой опасности? По мнению авторов, это:

— строгое соблюдение стандарта исключений для роботов (этот документ можно найти, например, по адресу http://www.robotstxt.org/wc/exclusion.html);

— аккуратное описание сценариев сбора информации роботами, зачастую буквально эмуляция действий пользователей;

— создание сети информационных прокси-серверов, например, на отраслевых уровнях. В этом случае сканироваться могут не web-сайты-оригиналы, а ближайшие прокси-серверы.

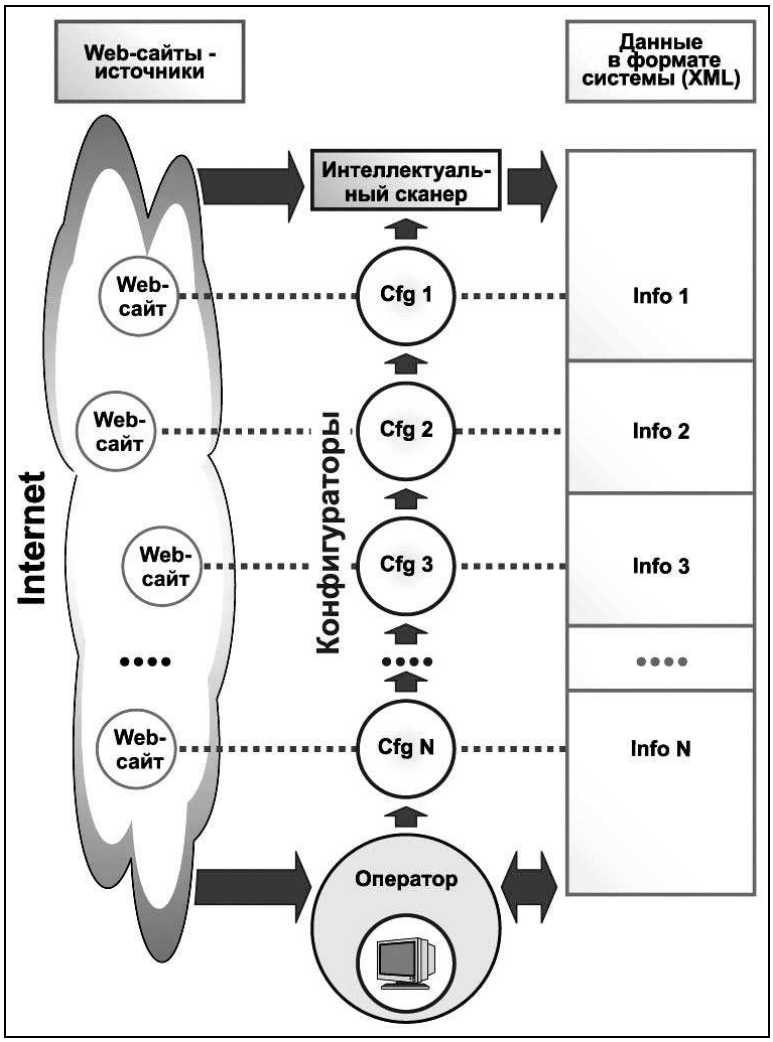

Рис. 1. Процедура сбора данных

На рис. 2 приведен принцип функционирования сети информационных прокси-серверов. На нем представлен иерархический принцип организации этой сети. Прокси-сервер первого уровня обеспечивает доступ к кэшу, заполняемому интеллектуальным сканером. К этому кэшу с помощью информационно-поисковой системы обеспечивается доступ конечных пользователей корпоративной сети. Эти же пользователи имеют возможность обращения к документам непосредственно в сети Интернет. Представленные на рис. 2 прокси-серверы 2-го уровня загружают информацию с кэша прокси-сервера 1-го уровня, а кроме того, могут дополнять свое информационное хранилище данными, сканируемыми непосредственно из

Интернет (информационные потребности пользователей разных прокси-серверов могут отличаться). Очевидно расширение приведенной схемы на третий и последующие уровни.

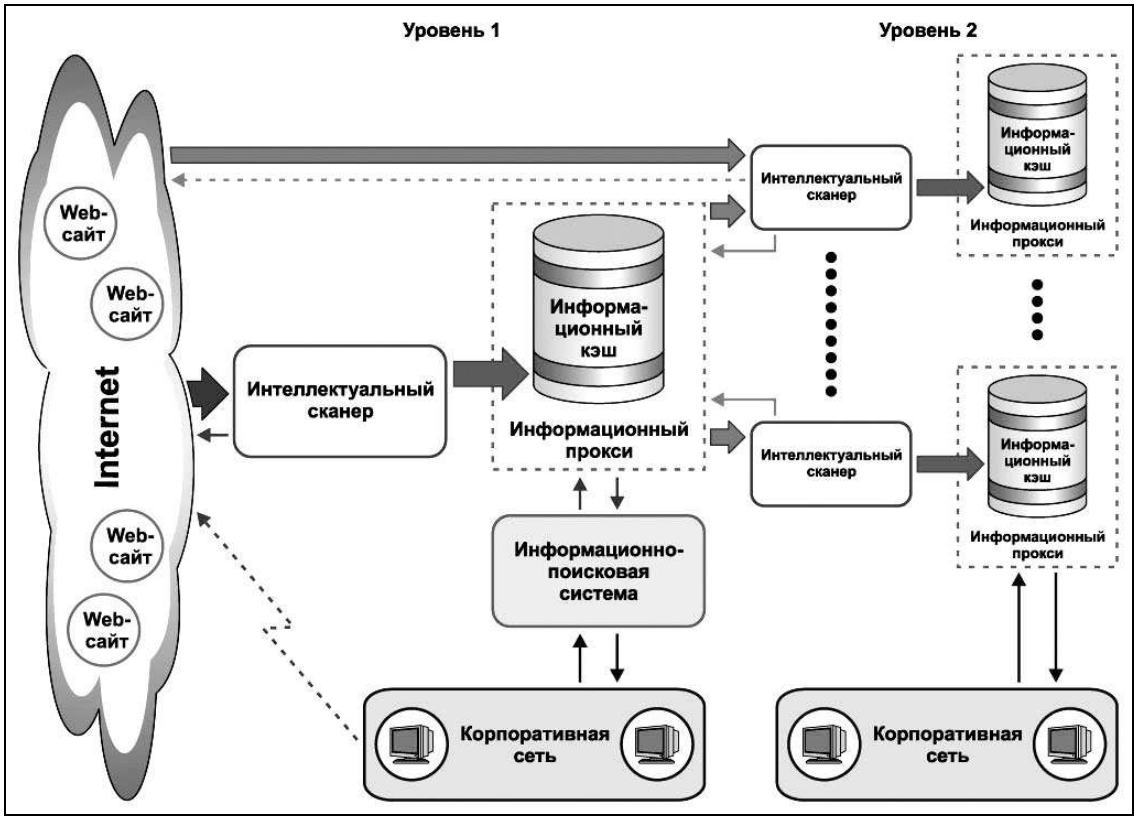

Рис. 2. Принцип организации сети информационных Proxy-серверов

Программно-аппаратный комплекс InfoStream Port

В качестве прототипа информационного прокси-сервера рассматривается система, созданная на основе комплекса мониторинга новостей InfoStream [9], которая в настоящее время позволяет осуществлять сканирование информации из нескольких тысяч открытых web-сайтов.

На основе этой системы реализуетися информационный прокси-сервер, к которому обращаются пользователи — корпоративные серверы, которые сами непосредственно не сканируют Интернет (или выполняют эту операцию в ограниченных объемах, решая специфические информационные задачи). Такой подход обладает следующими преимуществами.

1. Не требуется сканирования и обработки данных из Интернет непосредственно (прежде всего — экономия на ресурсах, необходимых для администрирования).

2. Анонимность (при сканировании сайтов их владельцы могут определять адреса робота-сканера).

3. Существенная экономия Интернет-трафика (в этом случае основные расходы берет на себя информационный провайдер — владелец первого проксисервера. Как показывает опыт, соотношение объемов сканируемой и «готовой к употреблению» информации составляет 50:1).

4. Не отрицается возможность самостоятельного сканирования Интернет (например, ресурсы общего плана можно загружать из информационного проксисервера, а специальные ресурсы — непосредственно из Интернет).

Для корпоративных пользователей реализовано решение InfoStream Port, которое обеспечивает доступ к базам данных оперативной и ретроспективной информации в корпоративных сетях. Программно-технологическое обеспечение InfoStream Port основано на принципе интеграции информационного проксисервера и поисковой системы и включает как компоненты утилиту обмена данными с информационным хранилищем (кэшем) и полнотекстовую информационно-поисковую систему InfoRes.

Информационное обеспечение системы у корпоративного пользователя, функционирование которой основывается на использовании кэша, формируется за счет выполнения совокупности технологических операций, в число которых входят сбор информации из Интернет, нормализация информации, приведение ее к единому системному формату, классификация, помещение данных в информационное хранилище и предоставление санкционированного доступа к кэшу.

Заключение

Описанная распределенная система информационных прокси-серверов позволяет создавать эффективные и масштабируемые решения, которые могут быть существенным подспорьем для аналитиков, сотрудников информационных служб, так как они способны существенно повысить надежность доставки и уровень обобщения оперативных данных, а также снизить загрузку каналов связи. Благодаря используемому кэшированию не только повышается эффективность использования каналов, но и уменьшаются задержки, возникающие в процессе доставки интернет-контента пользователю.

Критически важным в этой технологии являются инструментальные средства, которые должны гарантировать безопасность, актуальность принимаемых и передаваемых данных, а также их целостность.

1. *Додонов А.Г., Клещев Н.Т., Клименко В.Г.* Анализ отраслевых вычислительных сетей. — Л.: Судостроение, 1990. — 256 с.

2. Вимоги до захисту інформації WEB-сторінки від несанкціонованого доступу. НД ТЗІ 2.5-010-03. — К.: ДСТСЗІ СБ України, 2003. — 20 с.

3. *Уэйнрайт П.* Apache для профессионалов. — М.: Лори, Wrox Press Ltd, 2001. — 474 с.

4. *Дюбуа П.* MySQL. — М.: ИД «Вильямс», 2004. — 1056 с.

5. *Фурашев В.Н., Ландэ Д.В., Григорьєв А.Н., Фурашев А.В.* Электронное информационное общество Украины: взгляд в настоящее и будущее // Академия правовых наук Украины. Научно-исследовательский центр правовой информатики. — К.: Инжиниринг, 2005. — 163 с.

6. *Азаров С.С., Додонов А.Г.* Информационные техологии: кибервойны, информационные войны и сетевые войны: Сб. науч. тр. Международной конференции «Информационные технологии и безопасность». Вып. 5. — К.: Национальная академия наук Украины, 2003. — С. 3–11.

7. *Ландэ Д.В.* Данные в кармане // CHIP/Украина. — 2002. — № 6. — С. 82–85.

8. *Ландэ Д.В.* Сканер системы контент-мониторинга InfoStream // Открытые информационные и компьютерные интегрированные технологии: Сб. науч. тр. Вып. 28. — Харьков: Аэрокосмический ун-т «ХАИ», 2005. — С. 53–58.

9. *Ланде Д.В., Фурашев В.М, Григор'єв О.М.* Програмно-апаратний комплекс інформаційної підтримки прийняття рішень: Науково-методичний посібник. — К.: ТОВ «Інжиніринг, 2006. — 48 с.

Поступила в редакцию 17.08.2006